Erschienen als: Steyer, Kathrin (2003): Korpus, Statistik, Kookkurrenz. Lässt sich Idiomatisches "berechnen"? In: Burger, Harald/Häcki, Buhofer, Annelies/Gréciano, Gertrud (Hrsg.): Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie. S. 33-46 - Baltmannsweiler: Schneider Hohengehren, 2003. (Phraseologie und Parömiologie 14)

Korpus, Statistik, Kookkurrenz. Lässt sich Idiomatisches „berechnen"?

Vorbemerkung

Die Phraseologie- und Idiomatikforschung ist mittlerweile zu einer elaborierten linguistischen Disziplin geworden, mit einem umfangreichen Reservoir an Klassifikationsmodellen und detaillierten Einzeluntersuchungen. Mehr und mehr rückt jedoch die Notwendigkeit ins Blickfeld, all diese Erkenntnisse und Modelle anhand des tatsächlichen Sprachgebrauchs zu überprüfen, weiterzuentwickeln oder möglicherweise auch in manchen Teilen zu revidieren. Neben Informantenbefragungen stellt die Analyse fertiger Sprachprodukte, die Analyse von großen Textmengen, einen zukunftsweisenden Weg dar. Eine zentrale Methode im Bereich der automatischen Korpusanalyse ist die statistische Kookkurrenzanalyse[1], mit der es möglich ist, Auffälligkeiten im Verhalten von Wörtern zueinander statistisch zu erkennen und zu systematisieren. Die Methode erlaubt eine Verifizierung des Geltungsbereiches usueller Wortverbindungen in einer neuen Dimension, die weit über die Möglichkeiten kompetenzbasierter Konstruktionen hinausgeht.[2] Dazu braucht man gro-

---

[1] Kookkurrenz verstehen wir als Oberbegriff für das statistisch signifikante Miteinandervorkommen von Textwörtern (tokens). Usuelle Kookkurrenzen sind in erster Instanz binäre Relationen zwischen autosemantischen Wortschatzelementen (Kollokationen). Dazu gehören auch all jene Wortverbindungen, die einen Mehrwortstatus aufweisen, also selbst als lexikalisch-semantische, grammatische und/oder pragmatische Einheiten anzusehen sind (z.B. Idiome, kommunikative Formeln, Funktionsverbgefüge usw.). Ich danke in diesem Zusammenhang Dimitrij Dobrovol'skij für seine sehr instruktiven Kommentare zu diesem Kookkurrenzansatz.

[2] Zum Konzept der usuellen Wortverbindungen vgl. u.a. Steyer 2000; Steyer 2003.

ße elektronische Korpora mit Massendaten[3], intelligente Werkzeuge, die den Menschen in die Lage versetzen, mit diesen Massendaten sinnvoll umzugehen, und Konzepte, die diese Methoden für die linguistische Forschung fruchtbar machen. Eine solche Methode, die COSMAS-Kookkurrenzanalyse, die am Institut für Deutsche Sprache in Mannheim entwickelt wurde, soll im Folgenden beschrieben werden. Wir wollen zeigen, dass sie nicht allein zur Selektion von binären Kollokationsrelationen aus elektronischen Textkorpora dient, sondern auch Rückschlüsse auf das Vorkommen und den Charakter von (idiomatischen) Mehrworteinheiten schlechthin liefert.

1. Einige Prämissen zur Philosophie der statistischen Kookkurrenzanalyse[4]

Der folgende Exkurs soll dazu dienen, die Möglichkeiten und Interpretationsspielräume, aber auch die Grenzen der statistischen Kookkurrenzanalyse verstehen und einordnen zu können. Dabei beziehen wir uns nur auf den Bereich der Aufbereitung und Analyse von wirklichen Massendaten aus Korpora und nicht auf den schon seit Jahrzehnten existierenden Bereich der Sprachstatistik im Allgemeinen.

Die Korpuslinguistik hat – vor allem im angelsächsischen Raum – bereits zu beachtlichen Ergebnissen geführt;[5] eine Vielzahl von Forschergruppen auf der ganzen Welt beschäftigt sich mit der Entwicklung automatischer Werkzeuge und Analysemethoden. Dies hat aber auf die klassische Phraseologie- und Idiomatikforschung bisher kaum Auswirkungen gehabt. Ein Grund dafür ist sicherlich, dass die Korpuslinguistik – betrachtet man diese Entwicklung aus wissenschaftshistorischer Perspektive – eine sehr junge Disziplin ist. Damit verbunden ist die zweite Schwierigkeit: Die Korpuslinguistik ist per se interdisziplinär; man kommt ohne computertechnologischen Hintergrund nicht aus. Nun gibt es traditionel-

---

[3] Die IDS-Korpora stellen mit derzeit über 1,7 Milliarden laufenden Wortformen die größte Sammlung elektronischer deutschsprachiger Texte weltweit dar. Ca. eine Milliarde sind öffentlich zugänglich und kostenlos recherchierbar (http://www.ids-mannheim.de/kt)

[4] Die folgenden Ausführungen sind auch ein Ergebnis vieler gemeinsamer Diskussionen mit Cyril Belica, dem Entwickler unseres Korpusanalysemoduls „Statistische Kollokationsanalyse und Clustering" (http://corpora.ids-mannheim.de/cosmas © 1995-2002 Institut für Deutsche Sprache, Mannheim). Zum Einsatz automatischer Selektionsverfahren von Mehrworteinheiten im Kontext sprachtheoretischer Fragestellungen, insbesondere des Lexikons, vgl. auch ausführlich Lemnitzer 1997.

[5] Vgl. u.a. Sinclair 1991.

lerweise Verständnis- und Kommunikationsschwierigkeiten gerade zwischen Sprachwissenschaftler und Computerexperten. Linguisten misstrauen der Statistik, die sich so oft nicht an ihre Sprachintuition und ihr Sprachwissen hält. Sie bezweifeln, dass sich natürliche Sprache quasi berechnen lässt. Computerlinguisten und Statistiker halten ihre entwickelten Methoden bereits für das Ergebnis – der Weg ist das Ziel –, und folgerichtig halten sie kompetenzbasierte Systematisierungen und Interpretationsversuche von Linguisten für nicht nötig, subjektiv oder gar verfälschend. Will man nun zu Resultaten gelangen, die einerseits über das einfache Belegsammeln bzw. die pure Verifizierung eigener Vorannahmen in den Korpora, andererseits aber auch über die Entwicklung automatischer Analysemethoden als reinen Selbstzweck hinausgehen, müssen beide Seiten Barrieren überwinden.

Der Computerexperte muss sich von dem Gedanken verabschieden, dass Automaten die Arbeit der Linguisten irgendwann einmal völlig ersetzen könnten, da der Rechner selbst in der Lage sein würde, Sprache in ihrer Komplexität zu erfassen und zu beschreiben. Er muss die Interpretationsleistung des Linguisten anerkennen und – im idealen Fall – diese für seine Weiterentwicklungen nutzen.

Der Linguist muss sich ebenso einiger Illusionen entledigen und folgende drei Kriterien in Frage stellen: Repräsentativität, Sprachintuition und Frequenz.

Er muss sich zunächst von der Idee der absoluten Repräsentativität eines Korpus verabschieden, eines repräsentativen Korpus, das ihm dazu verhilft, **sprachlichen Usus** schlechthin beschreiben zu können. Repräsentativität kann in unserem Verständnis nur ein relationaler Terminus sein, ein Korpus ist repräsentativ in Bezug auf etwas (etwa auf einen Teil der Sprachgemeinschaft, auf einen Text- bzw. Diskursbereich, auf eine historische Etappe usw.). In diesem Sinne sind beispielsweise die IDS-Korpora nicht in erster Instanz repräsentativ, da sie sowohl nach quantitativen Kriterien als auch nach einer größtmöglichen Variabilität der externen Merkmale akquiriert und aufbereitet werden (stetig wachsende Massendaten, größtmögliche Textsortenvielfalt, diachrone Vielfalt usw.). Repräsentativität lässt sich aber doch herstellen, wenn man sich, wie in den IDS-Korpora möglich, eigene virtuelle Korpora zusammenstellt, die auf das konkrete Analyseinteresse bezogen sind, und diese systematisch auswertet. Deshalb sprechen wir von einer korpusbezogenen Usualität. Dementsprechend sind also auch Wortverbindungen immer usuell in Bezug auf das zu Grunde liegende Korpus. Je umfassender ein Korpus ist,

und zwar quantitativ und qualitativ, desto näher kann man dem sprachlichen Usus kommen.

Der Linguist muss des Weiteren seiner Sprachintuition misstrauen. Er wird bei korpusbasierten Kookkurrenzanalysen mit einem Spektrum sehr verschiedener Befunde konfrontiert. Er findet Kookkurrenzpartner vor, die einer assoziativen – also mit der eigenen Sprachintuition korrespondierenden – Zuordnung in besonderer Weise entsprechen. Diese Kookkurrenzpartner werden als prototypisch empfunden. Andere dagegen bewertet der Linguist eher als untypisch oder gar falsch. Die als falsch oder irritierend empfundenen oder zumindest in Frage gestellten Befunde führen uns zu den Prozessen im Hintergrund: Wir müssen uns nämlich eingestehen, dass unsere eigene Sprachkompetenz etwas sehr Subjektives und von vielen Faktoren Beeinflusstes ist. Wir haben ganz unterschiedliche Rezeptionswelten, rekapitulieren einmal Gelerntes immer und immer wieder, wir vergessen Wichtiges usw. Wir bewerten demzufolge auch die empirischen Befunde in der Regel ganz unweigerlich vor der Folie unserer individuellen Sprachgenese, zumeist im Gegensatzpaar: [„Das ist typisch"] vs. [„Das ist untypisch"] oder ["Habe ich erwartet"] vs. ["Habe ich nicht erwartet"]. Diese Perspektive muss zwangsläufig zu Irritationen und Missverständnissen führen, wenn sie auf die Hypothesen und Annahmen der statistischen Kookkurrenzanalyse trifft. Es entstehen Kollisionen zwischen zwei unterschiedlichen Prämissen: zwischen dem linguistischen Postulat der **sprachintuitiven Erwartbarkeit** einerseits und dem statistischen Prinzip der unterstellten **Unerwartbarkeit** des Miteinandervorkommens von Wörtern andererseits. Was bedeutet das? 'Erwartbar im linguistischen Sinne' ist gemeint als ein Miteinandervorkommen von Wörtern, das für den analysierenden Betrachter nachvollziehbar ist. Er geht nicht voraussetzungslos an Sprache heran. Er macht Vorannahmen auf der Basis seines sprachlichen Wissens und seiner Spracherfahrung. Er kennt Strukturen und hat auch sehr grobe Häufigkeitsvorstellungen; er hat eine intuitive Ahnung davon, was "normaler Sprachgebrauch" und damit erwartbar ist. Das Prinzip der statistischen Analyse ist jedoch anders: Der Rechner geht zunächst völlig voraussetzungslos und ohne Vorannahmen über mögliche Strukturen an die Analyse.[6] Seine Ausgangshypothese ist die unterstellte Unabhängigkeit von sprachlichen Zeichen. Allgemeiner formuliert: Der Beobachtungsgegenstand, das Korpus, wird zunächst nur als eine unstrukturierte und unspezifische Folge von

[6] Es ist klar, dass der Rechner nur das tut, was der Mensch ihm vorgegeben hat. Wir verwenden dieses Bild eines agierenden, denkenden und handelnden Rechners nur zur Veranschaulichung.

Zeichenketten angesehen, deren Elemente beziehungslos und zufällig zueinander existieren (statistisches Zufallsprinzip). Der Rechner geht weiter davon aus, dass nur ganz wenige Wörter wirklich häufig sind. Die statistische Analyse dient nun dazu, herauszufinden, ob es nicht doch innerhalb dieses angenommenen beliebigen Vorkommens von sprachlichen Zeichen Auffälligkeiten gibt, die es erlauben, von geordneten Strukturen, von Mustern innerhalb dieser ‚ungeordneten Welt' zu sprechen. Mit **auffällig** ist also in diesem Kontext 'statistisch überproportional häufig' gemeint. Der Rechner erfasst zunächst Strukturen, dann analysiert und schließlich ordnet er diese. Der Gegenstand der statistischen Beobachtung ist in unserem konkreten Fall das Verhalten von Wörtern in einer speziellen Umgebung, das gemeinsame Auftreten und die Anordnungsbeziehung von Zeichenketten in einem Sprachausschnitt (Kookkurrenz). Die Annahme einer rein zufälligen Wortverbindung kann verworfen werden. Bei der statistischen Kookkurrenzanalyse geht es also um die Aufdekkung von Kombinationen von Wörtern, die erstens statistisch **unerwartet** häufig nahe nebeneinander vorkommen – häufiger, als man nach dem Zufallsprinzip erwarten würde – und die zweitens häufiger genau an einer bestimmten Stelle auftreten, als dies in Bezug auf das Gesamtvorkommen der Wörter im Korpus zu erwarten gewesen wäre.

Eng mit der Erwartbarkeitshypothese hängt schließlich ein weiterer sehr verbreiteter Irrtum zusammen: die Annahme, dass die Selektion von usuellen Wortverbindungen nach dem Frequenzkriterium erfolgt. Bei usuellen Kookkurrenzen handelt es sich nicht in jedem Fall um hochfrequente Verbindungen. Im Gegenteil: Idiomatische Verbindungen weisen in der Regel keine hohe Frequenz auf, werden aber mit unserer Methode ebenso erfasst. Oft findet man idiomatische Kookkurrenzen sogar in den oberen Rängen der statistischen Signifikanzlisten. Der Rechner interessiert sich, wie eben beschrieben, für jegliche Auffälligkeiten in der Umgebung eines Wortes. **Auffällig** kann aber auch bedeuten, dass ein sehr seltenes Wort (z.B. eine unikale Einheit wie *balbieren*) immer in der Umgebung eines anderen Wortes (*Löffel*) vorkommt. Damit weist diese Wortverbindung trotz geringer Vorkommenshäufigkeit einen hohen Kohäsionsgrad auf. Auffällig kann ebenso bedeuten, dass in der Umgebung einer Kollokation weitere Wörter überproportional häufig vorkommen, dass also z.B. das Adjektiv *fest* eine Signifikanz in der Umgebung der Kollokation *Dach - Kopf* aufweist. Der Linguist kann in aller Regel nicht in die Black-Box des Rechners blicken. Er ist mit statistischen Listen konfrontiert, nicht mit den Prozessen im Hintergrund.

Nun ist es an der Zeit, den eingangs postulierten Widerspruch zwischen traditionell linguistischem Herangehen und korpusstatistischer Analyse zu relativieren. In den allermeisten Fällen erhält man bei der Analyse – trotz der bereits erwähnten überraschenden Ergebnisse – Resultate, die mit der eigenen Erwartungshaltung doch in Einklang zu bringen sind. Hier stellt sich dann der Widerspruch zwischen kompetenzbasierter Erwartbarkeit und statistischer Unerwartbarkeit als künstlich, oder besser als rein für heuristische Zwecke konstruiert heraus: Der Rechner hat eine Hypothese aufgestellt (die der ungeordneten Strukturen), um sie dann zu verwerfen und geordnete Zusammenhänge aufzudecken – solche Zusammenhänge, die auf usuellen Sprachgebrauch hindeuten.

2. Die COSMAS-Kookkurrenzanalyse[7]

Die COSMAS-Kookkurrenzanalyse ermöglicht kontextsensitive Extraktionen usueller Kookkurrenzen aus großen Korpora. Es handelt sich um ein interaktives und dynamisches Werkzeug, das sich im Gegensatz zu den meisten Tools dieser Art flexibel an unterschiedliche Korpusgegebenheiten anpassen kann und online auf dem neuesten Stand abrufbar ist. Der Ausgangspunkt der statistischen Kookkurrenzanalyse ist die Analyse des kookkurrenziellen Potenzials eines Einwortlemmas, und zwar anhand sehr unterschiedlicher Parameter und abgestufter Tiefen der statistischen Berechnung. Der Rechner analysiert die KWIC-Zeilen (Keyword-in-Context) eines Bezugswortes und sucht in diesem Ausschnitt nach Auffälligkeiten des Verhaltens von Wörtern in Bezug auf das Suchwort. Er schafft damit eine innere Struktur und Hierarchie der Belegmenge des Sprachausschnitts, die durch statistisch geordnete Listen von signifikanten Kookkurrenzpartnern eines Wortes sichtbar wird. Er nimmt Präferenzsetzungen vor. Die Relationen zwischen diesen Kookkurrenzpartnern und dem Bezugswort weisen völlig unterschiedliche Merkmale, Bindungsqualitäten und Spezifika auf. Die Kookkurrenzanalyse ermöglicht nicht nur die Erfassung binärer Wortrelationen (Kollokationen), sondern auch die Ermittlung usueller phrasaler Muster bis hin zu komplexen (idiomatischen) Mehrwortverbindungen und ihrer Kontexte. Gerade in Bezug auf die kontextsensitive Recherche und das Erkennen phrasaler Muster und Festigkeiten geht die COSMAS-

[7] Die COSMAS-Kookkurrenzanalyse verbindet verschiedene statistische Ansätze miteinander; die Berechnung der Signifikanzwerte erfolgt mit Hilfe des log-likelihood ratio (vgl. u.a. Dunning 1993). Zur konkreten Funktionsweise der COSMAS-Kookkurrenzanalyse vgl. Steyer 2003 und Belica 2002.

Kookkurrenzanalyse über die meisten Werkzeuge dieser Art hinaus. Ein Beispiel:

Suchwort &Leben

Analyse-Kontext:	5 Wörter links, 5 Wörter rechts, höchstens 1 Satz
Granularität:	grob
Zuverlässigkeit:	hoch
Clusterzuordnung:	eindeutig
Lemmatisierung:	ja
Funktionswörter:	ignoriert

Kollokatoren sind sortiert absteigend nach der berechneten Stärke der lexikalischen Kohäsion.

BelegNr	Gamma	Kollokatoren	Häufigkeit
1+273:	1262	rufen	273
274+159:	710	Mensch kommen	159
433+255:		Mensch	255
688+68:	601	gerufen	68
756+401:	521	kommen	401
1157+169:	515	Tod	169
1326+103:	397	retten	103
1429+60:	315	erwecken	60
1489+32:	235	ungeboren	32
1521+45:	216	Leib	45
1566+36:	208	Sterbe	36
1602+67:	198	kulturell	67
1669+51:	177	gesellschaftlich	51
1720+23:	177	einhauchen	23
1743+90:	174	lieb	90
1833+35:	166	alltäglich	35
1868+228:	164	lang	228
2096+64:	161	normal	64
2160+60:	161	menschlich	60
2220+94:	148	wirklich	94
2314+20:	142	lebenswert	20
2334+109:	123	öffentlich	109
2443+16:	118	selbstbestimmt	16
2459+103:	113	schwer	103
2562+184:	112	ganz	184

[...]

Zu jeder Kollokationsrelation sind KWICs (Keyword-in-Context) und die dazugehörigen Volltextbelege abrufbar:

KWIC

```
T95 chlose mit Scheinanmeldungen.  "Das wirkliche Leben ist anders als das Melde
T95 trömt. Und doch, ganz anders als im wirklichen Leben, lassen sich die Frauen
T95                  Aber anders als im wirklichen Leben steht hier niemand hüst
T97 schste Stadt ganz Spaniens. Doch im wirklichen Leben sieht es anders aus. Da
T97                  Aber anders als im wirklichen Leben muß Aschenputtel
```

> Wer erinnert nicht Theaterstücke, in denen Männer wie Archetypen über die Bühne holzen, stürzen, torkeln? Immerzu brüllt es aus ihnen, ihre Triebe treiben sie an, trinken können sie nie, ohne daß ihnen der Wein über Rock und Schoß strömt. Und doch, ganz anders als im **wirklichen Leben**, lassen sich die Frauen auf der Bühne nicht davon abhalten, diese Kotzbrocken leidenschaftlich zu lieben. "Ist das deutsch?" fragten beim letzten Berliner Theatertreffen im Mai 1994 US-amerikanische TheaterkritikerInnen, für die ich bei den Aufführungen grob übersetzen mußte. Wie sollte ich ihnen diese Frage beantworten? T95/JUN.26609 die tageszeitung, 22.06.1995, S. 13

Die derzeit umfassendste Anwendung erfährt diese Methode im IDS-Projekt „Wissen über Wörter", dem Hypertextinformationssystem zum deutschen Wortschatz.[8] Es handelt sich um einen integrativen Ansatz, bei dem ein komplexes automatisches Tool mit einem linguistischen Modell verbunden wird. Die Herangehensweise unterscheidet sich von den meisten anderen korpusbasierten Ansätzen dadurch, dass nicht spezielle Wortkombinationen ausgewählt und auf ihr Vorkommen im Korpus überprüft werden. Folgerichtig formulieren wir auch keine rein kompetenzbasierten Ausschlussbedingungen bzw. Restriktionen für das Kombinationsverhalten eines Wortes oder machen Annahmen über die Typik bzw. Usualität einer Wortverbindung. Wir analysieren vielmehr Sprachausschnitte in großer Dimension, um signifikante Wortkombinationen und ihre kontextuellen Einbettungen herauszufinden und sie dann linguistisch zu beschreiben, zu systematisieren und zu klassifizieren. Als Basis dient uns in der Pilotphase die COSMAS-Kookkurrenz-Datenbank, die das Kookkurrenzpotenzial von 1000 Lemmata vollständig aufbereitet hat. Diese 1000 Lemmata sind aus den 2000 frequentesten Wörtern der IDS-Korpora (unter Ausschluss der Artikel) nach inhaltlichen Kriterien ausgewählt worden. Damit ist es uns zum ersten Mal möglich, einen **systematischen** Sprachausschnitt mit unserer Methode zu analysieren. In einem nächsten Schritt werden diese Kookkurrenzpartner nach bestimmten Gruppen geordnet, um dann wiederum mit Hilfe der kontextsensitiven automatischen Analyse, aber auch mit Hilfe unserer linguistischen Kompetenz all jene Kookkurrenzpartner herauszufinden, denen wir eine Indikatorenfunktion für eine weitergehende Festigkeit zuschreiben können. Diese Wortverbindungen werden zu Kandidaten für ein **Mehrwortlemma**. Natürlich kommen dabei Kriterien und Ansätze der bisherigen Idiom- und Phraseologieforschung zur Anwendung. Es gibt aber auch Fälle, bei denen wir darüber nachdenken müssen, warum diese Wortkombinationen ein signifikantes Miteinandervorkommen aufweisen, mit dem bisherigen Beschreibungsapparat jedoch nicht zu

[8] Informationen zu diesem Vorhaben kann man auf der Projekthomepage http://www.ids-mannheim.de/wiw finden.

erfassen sind. Es sollte schon deutlich geworden sein, dass linguistische Kompetenz und Interpretation nach wie vor unabdingbar sind.

3. Verifizierungsmöglichkeiten im Bereich der Mehrwortverbindungen

Im Folgenden beschreiben wir Verifizierungsmöglichkeiten im Bereich der Mehrwortverbindungen, die mit Hilfe der empirischen Methode der statistischen Kookkurrenzanalyse zu erzielen sind. Als Beispiele haben wir die Kookkurrenzfelder der Nomina *Leben* und *Erde* ausgewählt.

3.1. Statuszuordnung

Betrachtet man zunächst nur die Kookkurrenzliste von *Leben*, lassen sich ohne Mühe lexikalische Einheiten finden, die auf eine Mehrwortverbindung hinweisen, z.B. *rufen, einhauchen, Leib*. Es werden aber auch Einheiten gefunden, die auf Grund ihrer Transparenz nicht auf eine weitergehende Festigkeit verweisen; Beispiele sind *retten, gesellschaftlich* oder *kulturell*. Diese Interpretation ist rein kompetenzbasiert. Man kann bei der Betrachtung einer ‚groben' Kookkurrenzliste, einer statistischen Liste, die noch keine oder wenige Kotextpartner mitliefert (wie zu &Leben in 2.), nur mit der eigenen Kompetenz entscheiden, ob eine Kollokation auch beispielsweise eine idiomatische Qualität besitzt. Viele Fügungen lassen sich jedoch auf dieser Basis nicht näher zuordnen. Dazu ist eine weitere Kotextspezifikation vonnöten. Diese Spezifikation kann nun auf verschiedene Weise erfolgen: Man kann die Kookkurrenzanalyse vertiefen, indem man eine immer feinere Granularität einstellt und nach bestimmten Clustern sucht, die mit einer Kollokation verbunden sind. Bei vielen Kollokationspartnern ist jedoch darüber hinaus die Analyse der KWIC-Zeilen (Keyword-In-Context) unabdingbar, um Aufklärung über einen eventuellen Einheitsstatus der Wortverbindung und damit über den Status dieser Verbindung als Mehrwortlemma zu erhalten.

So könnte man meinen, dass auch eine Kollokation wie *wirkliches Leben* eine freie (!) Wortverbindung ist bzw. die Bedeutung der Wortverbindung aus den Bedeutungen ihrer Elemente, des Adjektivs *wirklich* und des Nomens *Leben*, abgeleitet werden kann. Die KWIC-Zeilen zeigen dann aber, dass diese Kollokation sehr häufig in feste, sich wiederholende syntaktische Konstruktionen eingebettet ist und damit Formelhaftigkeit aufweist.

anders als im wirklichen Leben

aus dem wirklichen Leben gegriffen
im wirklichen Leben

Ein Vorteil der Analysemethode besteht also im Erkennen von formelhaften, nichtidiomatischen Fügungen, die banal wirken, aber auf Grund ihrer hohen Signifikanz tatsächlich Halbfertigprodukte der Sprache im Hausmannschen Sinne sind [9].

Unsere Methode kann – und das wäre der dritte Weg einer Kotextspezifizierung – auch dazu eingesetzt werden, Kollokationen einer nochmaligen Analyse zu unterziehen. Der Rechner sucht wiederum auf der Basis der KWIC-Zeilen – und nun der KWIC-Zeilen der jeweiligen Kollokation – nach weiteren Auffälligkeiten in der Umgebung dieser Wortverbindung, nach lexikalischen Einheiten, die überproportional häufig in der Nähe dieser Kollokation auftreten. Analysiert man beispielsweise die Kollokation *wirkliches Leben*, erhält man als einen signifikanten Kookkurrenzpartner die Partikel *eben*. Mit der Verwendung dieser Partikel verstärkt der Sprecher seine Bewertung eines Sachverhalts.[10]

```
eben

B00 agt sie. Aber Kerstin sieht eben im wirklichen Leben so aus wie in Big Broth
T92 k aufgetragen, eben wie manchmal im wirklichen Leben. Das Schnürschuh-Theate
T94                ... aber eben aus dem wirklichen Leben.
T99  auch einsetzen." Alles eben wie im wirklichen Leben, oder?
Z96 rauch- und keimfrei, steril wie das wirkliche Leben eben, lichtet sich der R
Z98 nvermeidbar. Es geht eben zu wie im wirklichen Leben
```

Eine weitere Gruppe sind Mehrwortverbindungen, die eine diskursive Relevanz im öffentlichen Diskurs besitzen. Auch hierfür finden sich in unseren Kookkurrenzlisten Kandidaten, die wir ebenso durch die Einbeziehung der KWICs oder gar der Volltextbelege entsprechend interpretieren können.

Schutz des ***ungeborenen*** *Lebens*
Recht auf Leben und körperliche ***Unversehrheit***

Schließlich erhalten wir durch die Möglichkeiten der kotextspezifizierenden Analyse auch Kookkurrenzpartner, die Basiselemente von Redewendungen, Sprichwörtern und Zitaten sind.

Es gibt kein ***richtiges*** *Leben im Falschen .*
Wer zu spät ***kommt****, den bestraft das Leben.*
Die ***Dinge*** *des Lebens*
Zum Leben zu wenig, zum ***Sterben*** *zu viel*

---

[9] Vgl. Hausmann 1985, S. 118.

[10] Zu kommunikativen Formeln vgl. Stein 1995.

3.2. Invarianz, Modifikationsresistenz, Modifikationsanfälligkeit

Die Kookkurrenz stellt auch eine wesentliche Hilfe bei der Verifizierung invarianter Strukturen einer Mehrwortverbindung und – damit verbunden – ihrer Basiselemente bzw. ihres Modifikationspotenzials dar. Das Problem der Normalform als subjektive Lexikografenabstraktion ist hinlänglich bekannt und diskutiert, ohne bisher nur annähernd gelöst zu sein.[11] Es geht auch hier immer um eine Invarianz in Relation zum Korpusvorkommen der Mehrwortverbindung. Unsere Pilotuntersuchungen haben dabei gezeigt, wie sehr unsere eigene Intuition gerade in Bezug auf solche invarianten Kerne täuschen kann.

Der invariante Kern der Kollokation *Leben-Leib* ist beispielweise durch die KWICs als Zwillingsformel *Leib und Leben* identifizierbar.

Leib

B98 die Polizei, obwohl Raser *Leib und Leben* anderer gefahrden. Klaut jemand fu
B00 uen können, dass der Staat *Leib und Leben*" schutze. "Gewaltbereitschaft und
B00 ht Gehör verschaffen, ohne *Leib und Leben* zu riskieren. Es sei deshalb falsc
M95 r bereit sei, Gefahren fur *Leib und Leben* in Kauf zu nehmen
M98 Ja, es bestunde Gefahr fur *Leib und Leben*, hieß es sogar. Im linken Hauschen
M00 Bahnen keine Gefahr fur *Leib und Leben* darstellen.
T88 es sechs Straftaten gegen *Leib und Leben*, 37 Diebstahle, drei Rauschgiftdel
T92 u leisten. Eine Gefahr fur *Leib und Leben* kann dort nicht unterstellt werden
T92 Angesichts neuer Gefahren fur *Leib und Leben* selbst zu schutzen, nicht zu unter

Gleichzeitig lässt sich das typische syntaktische und lexikalische Verwendungsmuster bestimmen: *Gefahr für Leib und Leben*. Es ist in den KWICs aber auch erkennbar, dass sowohl Nomina und Präpositionen, die mit *Leib und Leben* verbunden sind, als auch die einbettende syntaktische Konstruktion fakultativ und damit ersetzbar sind.

3.3. Kontexte und Verwendungsspezifik

Besonders ertragreich ist die empirische Methode in jenen Fällen, bei denen sich die Ursache für die Signifikanz eines Kookkurrenzpartners mit der eigenen Sprachkompetenz nicht erklären lässt.

So ist ein signifikanter Kollokationspartner des Bezugswortes *Erde* das Nomen *Scheibe*. Natürlich lässt sich recht schnell ein Zusammenhang zwischen *Erde* und *Scheibe* herstellen. Diese Kollokation wird, so sollte man annehmen, im Kontext astronomischer und historischer Diskurse als Zitat verwendet. Dies ist aber noch kein hinreichender Grund, sie als gegenwärtig usuelle Mehrwortverbindung zu interpretieren. Die KWIC-Zeilen zei-

[11] Vgl. hierzu die Problematisierungen bei Dobrovol'skij 1993; Burger 1998.

gen dann, dass *Erde* und *Scheibe* prototypischerweise als satzwertige Phrase *Die Erde ist eine Scheibe* verwendet werden. Es wird außerdem sichtbar, dass der angenommene thematische Zusammenhang zwar vorhanden ist, darüber hinaus aber diese Phrase in ganz unterschiedliche, thematisch nicht spezifizierte Kontexte eingebettet ist. Dies legt die Vermutung nahe, dass es sich um eine Redewendung handeln muss, die sich von ihrem Ursprung entfernt hat, was sich bei der detaillierten Analyse der KWICs und der dazugehörigen Volltextbelege bestätigt. Selbst diejenigen Verwendungsweisen, die sich im ursprünglichen Referenzrahmen des historischen Astronomiediskurses befinden, sind zumeist metaphorischer Natur.

Erde Scheibe

B97/710	als die Welt noch in Ordnung und **die Erde eine** muhelos uberschaubare **Scheibe**
B98/810	warf die Sonne einen Schatten. Wenn **die Erde eine Scheibe** ware, wie angenommen w
B98/812	Umschreibungen für die Redewendung "**Die Erde ist eine Scheibe**".
B00/010	nd die besseren Autofahrer und **dass die Erde keine Scheibe ist**, das weiß ich auc
	fruher, in etwa zu der Zeit, **als die Erde noch eine Scheibe war**, bewegten sic
B00/012	eorie sein. **Und in Wirklichkeit ist die Erde** doch **eine Scheibe**, John Kennedy wur
M89/909	Demokratie in Ungarn seien und **die "die Erde immer noch fur eine Scheibe halten"**
T90/NOV	" oder "**Der Sozialismus ist tot und die Erde eine Scheibe**". Ein Munchner Bundesw
T97/SEP	ie Renten sind sicher. **Und die Erde ist eine Scheibe!**"
Z96/607	**Daß die Erde keine Scheibe ist**, hat Rom Galileo
B94/H09	vor vielen, vielen Jahren, **als die Erde noch eine Scheibe war**, in einer Stu

Mit dieser Mehrwortverbindung drücken Sprecher auf ironische Weise aus, dass sie etwas für genauso unwahr, abwegig, aussichtslos, unglaubwürdig oder aber altmodisch und überholt halten wie die Aussage *Die Erde ist eine Scheibe*. Die folgenden Belege illustrieren diese Verwendungsweise besonders deutlich:

> Aber dann ein Satz wie "Glauben Sie doch einfach, dass die **Erde** eine **Scheibe** ist und dass Frauen Auto fahren können!" Mannheimer Morgen, 11.03.2000

> Schon die mit 0:1 verlorene Partie gegen Rumänien hatte mit 22 Millionen eingeschalteten Geräten einen neuen Soccer- Rekord erzielt. Das heutige Match, das ABC vom Kabelsender ESPN übernommen hat, um eine größere Reichweite zu gewährleisten, wird neue Maßstäbe setzen. Da macht es auch nichts, daß, wie böse Zungen bemerken, die Chancen, daß die **Erde** eine **Scheibe** ist, größer sind als die des US-Teams. die tageszeitung, 04.07.1994

> So betreibe Gauweiler "Realitätsverleugnung", wenn er per Volksbegehren den Satz "Bayern ist kein Einwanderungsland" in die bayerische Verfassung schreiben lassen wolle: "Genauso gut könnte er beschließen lassen, daß die **Erde** eine **Scheibe** ist." Süddeutsche Zeitung, 28.02.1998

Eine weitere Verwendungsnuance findet sich in Belegen wie den folgenden:

Wenn schon, dann hätte der Arme es wohl besser bei einem «normalen Bürger» probiert, wahrscheinlich wäre die Strafe nicht so hoch ausgefallen. Hoffen wir, dass er wenigstens mit seinem Anwalt mehr Glück hat. Wenn das Gerechtigkeit ist, ist wohl die **Erde** doch eine **Scheibe**! St. Galler Tagblatt, 28.08.1999

Hier drückt der Sprecher seine Enttäuschung darüber aus, dass etwas, das von ihm für abwegig bzw. nicht vorstellbar gehalten wurde, doch zu- bzw. eintrifft.

Mit dieser Wendung werden auch Menschen bezeichnet, die die Zeichen der Zeit nicht erkennen („von-gestern" sind).

In der Gründungserklärung der Bewegung hieß es, die Gruppe solle bei der "friedlichen Auflösung des diktatorischen Systems" helfen und den Zusammenbruch der Wirtschaft und das politische Chaos abwenden. Pozsgay sagte nach seiner Wahl, die Bewegung stehe Anhängern unterschiedlicher politischer Überzeugungen offen, nicht aber Leuten, die gegen die Entwicklung einer Mehrparteien-Demokratie in Ungarn seien und die "die **Erde** immer noch für eine **Scheibe** halten". Mannheimer Morgen, 18.09.1989,

Bestätigt wurden aber auch die Eindrücke Rühmkorfs, die dieser zwanzig Jahre zuvor in den "Jahren die Ihr kennt" (1972) notiert hatte. Im November 1990 diagnostiziert Rühmkorf (in "Tabu I") nach einem Besuch bei Grass, als dieser sich über einen ihm unangenehmen Zeitungsartikel beklagt hatte: "Literarische Präpotenzvorstellung aus einer Zeit, als die Welt noch in Ordnung und die **Erde** eine mühelos überschaubare **Scheibe** war." Berliner Zeitung, 16.10.1997

4. Schlussbemerkung

Wir möchten noch einmal betonen, dass man neuen empirischen Methoden nicht blindlings vertrauen sollte und sie – wie gezeigt – eine linguistische Analyse und Interpretation nicht ersetzen können. Sprache vermittelt sich nach wie vor und vor allem über Kommunikation und lässt sich in diesem Sinne nicht vollständig berechnen. Wir sollten aber die Möglichkeiten und Methoden, die uns die computerbasierte Korpusanalyse bietet, kennen und nutzen. Wir sollten sie als Chance verstehen, sprachlichen Usus in einer völlig neuen Dimension beschreiben zu können. Es sollte kein Gegensatz zwischen den Paradigmen aufgebaut, sondern eine produktive Synthese gefunden werden: Textkorpora **und** Lexikografie. Statistik **und** Pragmatik. Computer **und** Kultur.

Literatur

Belica, Cyril (2002): Die COSMAS-Kollokationsanalyse (http://www.ids-mannheim.de/kt/ kollok.html).

Burger, Harald (1998): Phraseologie. Eine Einführung am Beispiel des Deutschen. (= Grundlagen der Germanistik; 36). Berlin.

Dobrovol'skij, Dmitrij (1993): Datenbank deutscher Idiome. Aufbauprinzipien und Einsatzmöglichkeiten. In: Földes, Csaba (Hg.): Germanistik und Deutschlehrerausbildung, Szeged/Wien 1993, 51-67.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, Vol 19, 1.

Haß-Zumkehr, Ulrike (2002): Das Wort in der Korpuslinguistik. Chancen und Probleme empirischer Lexikologie. In: Agel, Vilmos /Gardt, Andreas/Haß-Zumkehr, Ulrike/Roelcke, Thorsten (Hg.): Das Wort ... Festschrift für Oskar Reichmann zum 65. Geburtstag. Tübingen. (erscheint)

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, H./Mugdan, J. (Hg.): Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984. (= Lexicographica 3). Tübingen, S. 118-129.

Lemnitzer, Lothar (1997): Akquisition komplexer Lexeme aus Textkorpora. (=Reihe Germanistische Linguistik, 180). Tübingen.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

Stein, Stephan (1995): Formelhafte Sprache. Untersuchungen zu ihren pragmatischen und kognitiven Funktionen im gegenwärtigen Deutsch. (= Sprache in der Gesellschaft. Beiträge zur Sprachwissenschaft; 22). Frankfurt a. M./Berlin/Bern/New York/Paris/Wien.

Steyer, Kathrin (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 2/00, 101-125.

Steyer, Kathrin (2002): Wenn der Schwanz mit dem Hund wedelt. Zum linguistischen Erklärungspotenzial der korpusbasierten Kookkurrenzanalyse. In: Haß-Zumkehr, U./Kallmeyer, W./Zifonun (Hg.): Ansichten zur deutschen Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag. (=Studien zur deutschen Sprache). Tübingen. (erscheint)

Steyer, Kathrin (2003): Idiomatik hypermedial. Zur Repräsentation von Wortverbindungen im Informationssystem „Wissen über Wörter". In: Palm Meister, Christine (Hg.): EUROPHRAS 2000. Akten der Internationalen Tagung zur Phraseologie 15.-18. Juni 2000 in Aske, Schweden.(= Stauffenburg Linguistik). Tübingen. (erscheint)